Improving Detection Timeliness by Modeling and Correcting for Data Availability Delays

Ben Y. Reis, Ph.D., Kenneth M. Mandl, M.D. M.P.H.
Children's Hospital Informatics Program at the Harvard-MIT Division of Health Science and Technology

OBJECTIVE

This paper describes an approach to improving the detection timeliness of real-time health surveillance systems by modeling and correcting for delays in data availability.

BACKGROUND

The performance of even the most advanced syndromic surveillance systems can be undermined if the monitored data is delayed before it arrives into the system. In such cases, an outbreak may be detected only after it is too late for appropriate public health response.

Surveillance systems can experience delays in data availability for a number of reasons: The process of transmitting data from data sources to the surveillance system can involve delays, especially in large systems where data is first aggregated across a national network of data sources before being transmitted to the surveillance system. Delays can also arise in the course of care, where, for example, a diagnosis is not available for a few days after the healthcare encounter.

It is important to minimize delays in data availability in order to maintain timeliness of detection [1]. When this is not possible, it is desirable to compensate for these data delays to minimize their effects.

METHODS

We propose a method of increasing detection timeliness of surveillance systems by modeling historical data availability patterns and then using these models to extrapolate final data counts based on early incomplete data reports.

We model the historical data availability patterns from various DOD healthcare facilities to the CDC Biosense surveillance system beginning in January, 2005. These models predict the percentage of total visits from a given date that will be available to the system on each of the seven days following that date. The model incorporates day-of-week effects and outputs both expected value and variance.

The stability of data availability patterns over time varies greatly across healthcare facilities. We evaluate an approach for accounting for this variability by adjusting the extrapolated counts by the variance of the data availability model.

Using a Poisson regression model with a 7-day exponential filter [2] and semi-synthetic outbreaks with a lognormal temporal distribution [3], we compare the outbreak detection performance of three approaches: 1. Monitoring the raw data; 2. Monitoring the extrapolated counts; 3. Monitoring the extrapolated counts adjusted by the variance of the availability models. For all three approaches, we set a benchmark specificity of 95%.

RESULTS

We find that using the raw counts (Approach 1) yields good detection sensitivity but poor timeliness, due to the delays in data availability. Using the extrapolated counts (Approach 2) improves the timeliness dramatically, but the detection sensitivity is reduced due to the variance of the extrapolation models. Using the adjusted extrapolated counts (Approach 3) yields a significant though slightly smaller improvement in timeliness than Approach 2, but is accompanied by a far smaller drop in sensitivity.

CONCLUSIONS

Of the models tested, the best tradeoff between timeliness and sensitivity was achieved by Approach 3 – monitoring the extrapolated counts adjusted by the variance of the data availability. These results suggest that surveillance systems experiencing delays in data availability may benefit from this approach to improving their detection timeliness.

Since monitoring the raw counts (Approach 1) will likely yield greater detection sensitivity, we recommend a hybrid system that combines Approaches 1 and 3, as follows: Approach 3 is used to detect outbreaks early, and Approach 1 is used as a backup to identify any outbreaks missed by Approach 1. By integrating the outputs of these two approaches in this fashion, a system can reap the benefits of both without the issue of duplicate alarms.

REFERENCES

[1] CDC, Early Event Detection component of the PHIN Preparedness specifications. www.cdc.gov/phin/preparedness/eed.html

[2] Reis, B.Y., Pagano, M. & Mandl, K.D. (2003) Using temporal context to improve biosurveillance. Proc Natl Acad Sci U S A 100, 1961-5.

[3] Sartwell, P.E (1950) The Distribution of Incubation Periods of Infectious Disease, Am. J. Hyg. 51, 310-318; reprinted in Am. J. Epidemiol. 141, 386-94. (1995).

Further Information: Ben Reis, reis@mit.edu